

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号  
特開2000-259359  
(P2000-259359A)

(43)公開日 平成12年9月22日(2000.9.22)

(51)Int.Cl. <sup>7</sup>	識別記号	F I	テーマコード(参考)
G 0 6 F 3/06	3 0 6	G 0 6 F 3/06	3 0 6 B 5 B 0 1 8
	3 0 5		3 0 5 C 5 B 0 6 5
	5 4 0		5 4 0 5 D 0 6 6
12/16	3 2 0	12/16	3 2 0 L
G 1 1 B 19/02	5 0 1	G 1 1 B 19/02	5 0 1 F

審査請求 未請求 請求項の数 7 O L (全 13 頁) 最終頁に続く

(21)出願番号 特願平11-56883  
(22)出願日 平成11年3月4日(1999.3.4)

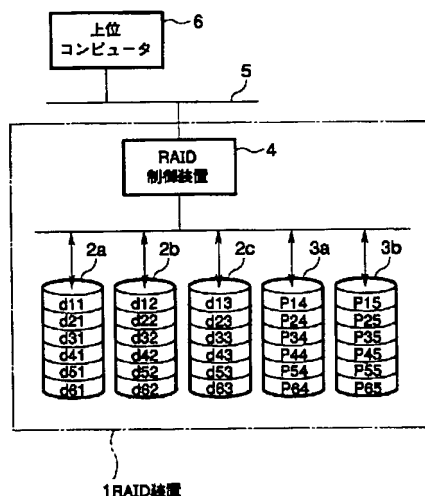
(71)出願人 000003078  
株式会社東芝  
神奈川県川崎市幸区堀川町72番地  
(72)発明者 伊藤 章  
東京都府中市東芝町1番地 株式会社東芝  
府中工場内  
(74)代理人 100058479  
弁理士 鈴江 武彦 (外6名)  
Fターム(参考) 5B018 GA02 HA16 KA21 MA14 QA16  
RA02  
5B065 BA01 CA30 EA02 EA12 EA24  
5D066 BA02 BA08

(54)【発明の名称】 RAID装置および記録媒体

(57)【要約】

【課題】 データディスクに冗長するパリティの算出およびディスク障害時のデータ復元を行うことにある。

【解決手段】 データを複数のデータ用ディスク2a～2cに分散保存し、これらデータに冗長するパリティ用ディスク3a、3bを付加する冗長手段と、データおよびパリティ保存領域をビット単位に任意整数にブロック分けするブロック分割手段と、2の拡大ガロア体GF(2n)を用いて、全ディスクのグループ単位ごとに前記データ用ディスクに付加するパリティ用ディスクのパリティを算出するパリティ算出手段および任意のデータ用ディスク障害時、障害発生ディスクのデータを未知データとし、前記拡大ガロア体GF(2n)で定める規則に従って連立合同式を作成し、この連立合同式から前記未知データを復元するデータ復元手段を有するRAID制御装置4とを設けたRAID装置である。



【特許請求の範囲】

【請求項1】 データを複数のデータ用ディスクに分散保存するRAID装置において、前記複数のデータ用ディスクに複数台のパリティ用ディスクを冗長する冗長手段と、前記データ用ディスクのデータ保存領域およびパリティ用ディスクのパリティ保存領域をビット単位に任意整数にブロック分けするブロック分割手段と、拡大ガロア体GF(2n)(nは整数)を用いて、前記全ディスクの所定ブロックどうして連なるグループ単位ごとに前記データ用ディスクに付加するパリティ用ディスクのパリティを算出するパリティ算出手段および前記任意のデータ用ディスク障害時、障害発生ディスクのデータを未知データとし、前記拡大ガロア体GF(2n)で定める規則に従って連立合同式を作成し、この連立合同式から前記未知データを復元するデータ復元手段を有するRAID制御装置とを備えたことを特徴とするRAID装置。

【請求項2】 請求項1に記載するRAID装置において、前記パリティを特定のディスクに固定せずに、前記グループ単位ごとに任意のディスクに分散するパリティ分散手段を設けたことを特徴とするRAID装置。

【請求項3】 複数のデータ用ディスクのデータを冗長するパリティを算出するパリティ算出プログラムを記録する記録媒体であって、コンピュータに、各データ用ディスクのグループ単位ごとのデータを読み出すデータ読み出し機能と、このグループ単位ごとのデータに対し、拡大ガロア体GF(2n)に従って所定の係数を乗算するとともに、これら乗算値を加算してパリティを算出するパリティ算出機能と、このパリティ算出機能によって算出されるパリティを該当グループ内の所定のブロックに設定するパリティ設定機能と、上記一連の機能を、全グループについて繰り返し実行する機能とを実現させるためにパリティ算出プログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項4】 データ用ディスク障害時にデータを復元するデータ復元プログラムを記録する記録媒体であって、コンピュータに、ディスク故障時に正常なデータ用ディスクのデータおよびパリティ用ディスクのパリティを読み出すデータ読み出し機能と、これら読み出したデータおよびパリティを用いて、拡大ガロア体GF(2n)の規則に基づいてディスク故障台数に応じた連立合同式を作成し未知データを求めるデータ復元機能と、これら一連の機能について全グループについて繰り返しデータを復元する機能とを実現させるためにデータ復元プログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項5】 請求項1または請求項2に記載するRAID装置において、

上位機器または自身のハードウェアスイッチから前記RAID制御装置に対して、冗長するパリティ算出の任意個数を決定し入力する手段を設けたことを特徴とするRAID装置。

【請求項6】 請求項1、請求項2および請求項5の何れか1つに記載のRAID装置において、それぞれ複数のディスクを管理する複数のRAID制御装置を連携させ、これらRAID制御装置管理下の全ディスクの所定ブロックをグループ化し、各グループごとにパリティを設定する手段を設けたことを特徴とするRAID装置。

【請求項7】 請求項1、請求項2、請求項5および請求項6の何れか1つに記載のRAID装置において、複数個のパリティを算出する場合、2個目以降のパリティについて、既に算出されたパリティをデータの一部とみなしてパリティを計算し、この計算結果のデータ誤りから障害ディスクを特定し、データを修復する手段を設けたことを特徴とするRAID装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ディスクの障害によって失われるデータを復元するRAID (Redundant Arrays of Inexpensive Disks) 装置および記録媒体に関する。

【0002】

【従来の技術】RAID装置は、各種のデータを複数の安価なディスクに分散し保存することにより、高性能ディスクと同等の性能を得る装置である。

【0003】しかし、小型のディスクを多数使用すれば、ディスクの故障が増加し、それだけデータが消失する危険性も高くなるので、単にディスクの数を増やすだけでなく、余分なディスクを用意し、冗長性をもたせることにより、安定性および高性能化を図っている。

【0004】RAID装置は、データ用ディスクの他に、冗長性をもたせるためのパリティ用ディスクが設けられ、データ用ディスクに障害が発生したとき、冗長性をもったパリティデータを用いて、障害によって失われたデータを復元することが行われている。

【0005】ところで、この種のRAID装置は、ディスクの組み合わせないし負荷分散の観点から、大まかには6つのレベルRAID1～RAID6に分類され、ディスク障害時のデータの復元化を図っている。

【0006】RAID1は、ミラーリング (mirroring) とも呼ばれる方式であって、この方式は、データ用ディスクが2つのグループに分けられ、同一のデータを2つのグループデータディスクに保存する二重化データ保存方式である。データの書込みは両グループ同時に行い、データの読み出しは何れかのデータ用ディスクから行う。その結果、一方のグループデータ用ディスクの障害時、別のグループに切替えるだけで消失データを簡単

に復元できる。

【0007】RAID2は、データを複数のディスクにビット単位で分散記録する一方、コンピュータのメモリで利用されているエラー訂正コード(ECC)を付加し、データ用ディスクの障害により失われたデータについて、エラー訂正コードを用いて復元し信頼性を高める方式である。

【0008】RAID3は、ディスクごとに障害を検知する仕組みがあることを前提とし、ECCを使わずにデータをビット単位で複数のディスクに分散させ、パリティ用ディスク1台を追加し、データ用ディスクの障害により失われたデータについて、ビット単位の排他的論理和によって障害データ用ディスク1台を復元する方式である。

【0009】RAID4は、ディスクごとに障害を検知する仕組みがあることを前提とし、RAID3がデータをビット単位で分散させたのに対し、データをブロック単位で複数のデータ用ディスクに分散させ、パリティ用ディスク1台を追加し、障害により失われたデータについて、ブロック単位の排他的論理和によって障害データ用ディスク1台を復元する方式である。

【0010】RAID5は、同じくディスクごとに障害を検知する仕組みがあることを前提とし、パリティをブロック単位に全てのディスクに分散保持させ、障害により失われたデータは、ブロック単位の排他的論理和によって障害データ用ディスク1台を復元する方式である。

【0011】RAID6は、RAID5が障害対策に単一パリティを用いたのに対し、パリティを2次元に拡張したり、Reed-Solomonコードを用いてエラー訂正機能を強化することにより、複数台のデータ用ディスク障害を復元し、信頼性を高める方式である。

【0012】

【発明が解決しようとする課題】従って、以上のようなRAID装置は、RAID1～RAID6に分類分けし、それぞれデータ用ディスク障害時のデータ復元を図っているが、各RAIDにはそれぞれ以下のような問題が指摘されている。

【0013】先ず、RAID1は必要な記憶容量の2倍のディスクを用意する必要があること、また当該RAID1を含めてECCを使用するRAID2やパリティを2次元的に拡張し設定するRAID6は、データ用ディスクとパリティ用ディスクとの台数比率が各方式ごとに固定となり、自由なデータ用ディスクとパリティ用ディスク台数設定ができないばかりか、一般的にはデータ用ディスクに対してパリティ用ディスクの占める割合が大きくなる問題があり、また少ない量のデータ変更がある場合であっても、多くのパリティを変更しなければならない問題がある。

【0014】さらに、RAID3～RAID5は、データ容量を増やすためにデータ用ディスクを追加すること

があるが、データ用ディスク追加にも拘わらず、パリティを排他的論理和にて計算する方式のものは、1台のデータ用ディスク容量分である1つのパリティしか設定できない。

【0015】その結果、データ用ディスクが増えたとき、同時に2台以上のデータ用ディスクの故障が発生する確率が高くなるばかりでなく、同時に2台以上のディスクが故障したとき、データが復元できなくなり、RAID装置の信頼性を損なう問題がある。

【0016】さらに、データ用ディスクを追加してデータ消失時のデータの復元を図る手段として、複数台単位ごとにグループ化し、各グループごとに1つのパリティを付け、信頼性を高めることが考えられる。

【0017】しかし、このような場合でも、同時に2台以上のデータ用ディスクが故障したとき、グループ内の故障ディスクが1台であれば復元可能であるが、同一グループ内において2台以上のデータ用ディスクに障害が発生したとき、失われたデータを復元できない問題がある。

【0018】本発明は上記事情にかんがみてなされたもので、データディスクに冗長するパリティを容易に算出し、また同時に複数台のディスクが故障した場合でも、そのディスク内容を復元可能なRAID装置を提供することにある。

【0019】また、本発明の他の目的は、用途等に応じて冗長するパリティの個数を容易に変更可能なRAID装置を提供することにある。

【0020】さらに、本発明の他の目的は、データ誤りディスクを特定し、そのデータの修復を可能とするRAID装置を提供することにある。

【0021】本発明の他の目的は、データディスクに冗長するパリティを算出し、また同時に複数台のディスクに故障が発生した場合でも、そのディスク内容を復元可能なプログラムを記録した記録媒体を提供することにある。

【0022】

【課題を解決するための手段】上記課題を解決するために、本発明は、複数のデータ用ディスクに複数台のパリティ用ディスクを冗長する冗長手段と、前記データ用ディスクのデータ保存領域およびパリティ用ディスクのパリティ保存領域をビット単位に任意整数にブロック分けするブロック分割手段と、拡大ガロア体GF(2n)

(nは整数)を用いて、前記全ディスクの所定ブロックどうして連なるグループ単位ごとに前記データ用ディスクに付加するパリティ用ディスクのパリティを算出するパリティ算出手段および前記任意のデータ用ディスク障害時、障害発生ディスクのデータを未知データとし、前記拡大ガロア体GF(2n)で定める規則に従って連立合同式を作成し、この連立合同式から前記未知データを復元するデータ復元手段を有するRAID制御装置とを

設けたRAID装置である。

【0023】なお、上記構成においては、パリティを特定のディスクに固定したが、グループ単位ごとに任意のディスクにパリティを分散する構成でもよい。

【0024】本発明は、以上のような手段を講じたことにより、複数のデータ用ディスクにデータが格納された後、パリティ算出手段にて、これら格納されたデータに基づき、拡大ガロア体GF(2<sup>n</sup>) (nは整数)を用いて、正確にパリティを算出しパリティ用ディスクに設定できる。

【0025】また、データ復元手段においては、データ用ディスク障害時、2の拡大ガロア体GF(2<sup>n</sup>)で定める規則に従い、正常ディスクデータ、パリティと障害発生ディスクのデータを未知データとして連立合同式を組み立て、この連立合同式から未知データを復元するので、複数台のデータディスクに障害が発生しても、迅速、かつ、確実に障害ディスクのデータを復元できる。

【0026】また、別の発明は、上位機器または自身のハードウェアスイッチからRAID制御装置に対して、冗長するパリティ算出の任意個数を決定し入力するようにすれば、用途、目的に応じて、冗長するパリティの個数を容易に変更可能であり、信頼性の向上およびユーザの要望に応じて柔軟に対処できる。

【0027】さらに、別の発明においては、それぞれ複数のディスクを管理する複数のRAID制御装置を連携させ、これらRAID制御装置管理下の全ディスクの所定ブロックをグループ化し、各グループごとにパリティを設定するようにすれば、パリティ用ディスクの台数を減らすことができ、RAID装置全体として冗長するパリティの低減化およびコストの削減に大きく貢献する。

【0028】さらに、別の発明は、複数個のパリティを算出する場合、2個目以降のパリティについて、既に算出されたパリティをデータの一部とみなしてパリティを計算することにより、この計算結果のデータ誤りから容易に障害ディスクを特定しデータ修復することができる。

【0029】

【発明の実施の形態】以下、本発明の実施の形態について図面を参照して説明する。

【0030】図1は本発明に係わるRAID装置の一実施の形態を説明する構成図である。

【0031】同図において1は例えば3台のデータ用ディスク2a、2b、2cおよび2台のパリティ用ディスク3a、3bを備えたRAID装置であって、これは伝送ライン5を介して上位コンピュータ6と接続されている。

【0032】このコンピュータ6は、例えばRAID装置1においてデータ処理に必要なプログラムなどを当該RAID装置1にダウンロードしたり、さらには当該RAID装置1に必要な指示、命令を与える機能をもって

いる。

【0033】前記RAID装置1は、上述したデータ用ディスク2a、2b、2cおよびパリティ用ディスク3a、3bの他、RAID制御装置4が設けられている。

【0034】前記各ディスク2a~2c、3a、3bは、4ビット単位でブロック分割され、かつ、1ディスク当たり6ブロックからなっている。

【0035】これらディスク2a~2c、3a、3bの各ブロックのうち、先頭(上位)ブロックどうしの5ブロックd11、d12、d13、p14、p15を1番目グループ、上位から2つ目のブロックどうしの5ブロックd21、d22、d23、p24、p25を2番目グループ、…と順次下位ブロックに対してグループ名が付けられている。

【0036】また、各ディスク2a~2c、3a、3bは4ビット単位で分割されて1ブロックとしているので、各ブロック内のデータは0~15までの16種類のデータが格納される。

【0037】前記RAID装置4は、外部から直接読み取り或いはコンピュータ6からダウンロードされたプログラムに従って、次のような手段を実行する。

【0038】すなわち、このRAID装置4は、2の拡大ガロア体GF(2<sup>n</sup>)を用いて、データ用ディスク2a~2cに付加されるパリティ用ディスク3a、3bのブロック単位に保存すべきパリティを算出するパリティ算出手段と、任意のディスク障害時、その障害発生ディスクのデータを未知データ(未知数)として2の拡大ガロア体GF(2<sup>n</sup>)の規則に従って連立合同式を組み立て、この組み立てられた連立合同式を解くことにより、障害発生ディスクのデータを復元するデータ復元手段を有するものである。

【0039】先ず、RAID装置4によるパリティの算出例について説明する。

【0040】各ディスクは4ビット単位でブロック分割されているので、各ブロックには0~15までの16種類のデータが格納可能である。

【0041】そこで、パリティの算出は、16種類のデータを扱える拡大ガロア体GF(extension Galois field)(2<sup>4</sup>)を用いて計算する。この拡大ガロア体は、定義された種類のデータ内で加算(減算)、乗算および除算を行うことができる。加算(減算)は排他的論理和を用いて計算し、乗算は図2に示す拡大ガロア体GF(2<sup>4</sup>)の乗算結果図を用いて計算する。なお、除算は図2の乗算結果図を逆引きすることにより引用する。

【0042】この図2の乗算結果を表す図は、GF(2)上の4次の規約多項式であるX<sup>4</sup>+X+1を用いてGF(2<sup>4</sup>)を計算した表である。

【0043】次に、簡単な拡大ガロア体の説明および図2の乗算結果を表す図の作成例について説明する。

【0044】ガロア体GF(2)は、0と1の2種類が

元であり、その間で加減乗除を行うことができる。これに対し、GF(2<sup>m</sup>)では2<sup>m</sup>個の元があり、その間で自由に加減乗除を行うことができる。2<sup>m</sup>は、2のm乗を表している。

【0045】ところで、拡大ガロア体における4次の乗算について定義する。

【0046】GF(2)の体となる0, 1の2種類の元において、図2の乗算結果であるGF(2)上の4次の規約多項式

$$X^4 + X + 1 = 0$$

の根を考えると、0, 1の何れの元を代入しても根が存

1	$\alpha$
2	$\alpha^2$
3	$\alpha^3$
4	$\alpha^4 = \alpha + 1$
5	$\alpha^5 = \alpha^2 + \alpha$
6	$\alpha^6 = \alpha^3 + \alpha^2$
7	$\alpha^7 = \alpha^4 + \alpha^3 = \alpha^3 + \alpha + 1$
8	$\alpha^8 = \alpha^4 + \alpha^2 + \alpha = \alpha^2 + 1$
9	$\alpha^9 = \alpha^3 + \alpha$
10	$\alpha^{10} = \alpha^4 + \alpha^2 = \alpha^2 + \alpha + 1$
11	$\alpha^{11} = \alpha^3 + \alpha^2 + \alpha$
12	$\alpha^{12} = \alpha^4 + \alpha^3 + \alpha^2 = \alpha^3 + \alpha^2 + \alpha + 1$
13	$\alpha^{13} = \alpha^4 + \alpha^3 + \alpha^2 + \alpha = \alpha^3 + \alpha^2 + 1$
14	$\alpha^{14} = \alpha^4 + \alpha^3 + \alpha = \alpha^3 + 1$
15	$\alpha^{15} = \alpha^4 + \alpha + 1$
16	$\alpha^{16} = \alpha$

以上の関係式から、当然 $\alpha$ のn次どうしを乗算しても、15種類の何れかの関係式の中に納まることになる。例えば、

$$\alpha^7 + \alpha^{14} = \alpha^{21} = \alpha^6 = \alpha^3 + \alpha^2$$

となり、15種類の中の式 $\alpha^6$ に納まる。この関係式からデータの種類「1」～「15」で一巡するので、 $\alpha^{16} = \alpha$ となり、「1」にもどる。

【0050】ここで、上記する15種類の関係式につき、1～15に対応させることにより、図2に示す乗算結果図を作成できる。この図2は $\alpha$ の指数を2進数の桁に対応させて数値化する。

【0051】因みに、上記例における $\alpha$ のn次の乗算は、以下のような計算として扱える。

$$【0052】11 \times 9 = 12$$

但し、本文中においては、通常の計算と区別する意味から、

$$11 \times 9 \equiv 12$$

のような記号で表す。なお、図2は、1～15に加えて0を含め、0～15の乗算結果を表す図となっている。

【0053】従って、図2は、GF(2)上の4次の規約多項式 $X^4 + X + 1$ を用いて、GF(2<sup>4</sup>)による乗算結果であるが、以下に述べる合同式についても、特にことわりがない限り、全てGF(2<sup>4</sup>)で計算している

在しない。そこで、このような規約多項式の根の一つを $\alpha$ と定義し体を拡大してみる。

【0047】その結果、 $\alpha$ は、 $\alpha^4 + \alpha + 1 = 0$ を満たすので、

$$\alpha^4 = \alpha + 1$$

の関係式が得られ、 $\alpha$ で表わされる4次以上の関係式は全て4次以下に置換することができる。

【0048】そこで、16種類のデータを $\alpha$ のn次で考えると、以下に示すような関係式で表すことができ、15種類の関係式以外には現れない。

【0049】

ものとする。

【0054】そこで、各ディスクd11, d12, d13の1番目グループに記録するデータに関し、1つ目のパリティp14の計算は、各データに1の係数を乗算した後に加算することにより求める。係数を1とするパリティの算出は従来のRAIDにおけるパリティ計算の場

合と同じである。具体的には、

$$1 \times d11 \oplus 1 \times d12 \oplus 1 \times d13 \equiv p14$$

$$d11 \perp d12 \perp d13 \equiv p14$$

となる。上式において記号 $\perp$ は排他的論理和で計算することを意味し、記号 $\times$ は図2の乗算結果を用いた乗算を意味し、記号 $\equiv$ は合同を意味する。

【0055】次に、各ディスクd11, d12, d13の1番目グループに記録するデータに関し、2つ目のパリティp15の計算は各データに任意の異なる係数を乗算し、この乗算により得られる値を加算することにより求める。

【0056】

$$2 \times d11 \oplus 3 \times d12 \oplus 4 \times d13 \equiv p15$$

今、具体例として、例えば各データd11, d12, d13が次のような値と仮定することにより、パリティを算出し、パリティブロックに設定する。

【0057】

$d_{11}=5$   
 $d_{12}=6$   
 $d_{13}=7$   
 $5 \perp 6 \perp 7 = p_{14}=4$   
 $2 \times 5 \perp 3 \times 6 \perp 4 \times 7 = p_{15}$   
 $10 \perp 10 \perp 15 = p_{15}=15$

以下、同様にして2番目グループ以降についても同様にパリティを算出し、パリティブロックに設定する。

【0058】次に、RAID制御装置によるデータ復元処理例について説明する。

【0059】今、全てのデータ用ディスク2a～2cが故障していない場合、またはパリティ用ディスク3a、3bが故障している場合、単にデータ用ディスクのデータを読み出すことにより、記録したデータを知ることができる。また、1台のみのデータ用ディスクが故障した場合には、従来のRAID方式と同じ排他的論理和により容易に障害ディスクのデータが分かる。

【0060】そこで、以降、2台同時期にデータ用ディスクが故障した場合のデータ復元例について説明する。

【0061】今、5台のディスク中2番目と3番目のディスクが同時期に故障し、1番目グループのデータd1

2、d13が失われたとする。この状態において1番目グループのパリティを含むd11、p14、p15は、ディスク2a、3a、3bが故障していないので、データの読み出しは可能である。

【0062】そこで、失われたデータd12、d13を未知データとして合同式の左辺に置き、読み出し可能な正常なデータを右辺に置き、合同式を作成すれば次の2つの合同式を作成することができる。

【0063】

$d_{12} \perp d_{13} = p_{14} \perp d_{11}$   
 $3 \times d_{12} \perp 4 \times d_{13} = p_{15} \perp 2 \times d_{11}$   
 このようにして得られる2つの合同式に対し、ディスクから正常なデータd11、p14、p15を読み出して式に代入し後、これら2つの連立合同式から未知データd12、d13を求める。この未知データを求める手順は、一般の連立方程式の解法と全く同様の手順で行う。そこで、各式に式番号( )を付加して説明を進めることとする。

【0064】

$$\begin{aligned}
 d_{12} \perp d_{13} &= 4 \perp 5 = 1 & \cdots (1) \\
 3 \times d_{12} \perp 4 \times d_{13} &= 15 \perp 2 \times 5 = 5 & \cdots (2)
 \end{aligned}$$

ここで、4×(1)式を計算し、下記する(3)式を得る。

$$\begin{aligned}
 4 \times d_{12} \perp 4 \times d_{13} &= 4 \times 1 \\
 4 \times d_{12} \perp 4 \times d_{13} &= 4 & \cdots (3)
 \end{aligned}$$

そこで、(3)式⊥(2)式を計算し、未知データd13を消去することにより、(4)式を作成する。

$$\begin{aligned}
 4 \times d_{12} \perp 3 \times d_{12} \perp 4 \times d_{13} \perp 4 \times d_{13} &= 4 \perp 5 \\
 (4 \perp 3) \times d_{12} \perp (4 \perp 4) \times d_{13} &= 4 \perp 5 \\
 7 \times d_{12} \perp 0 \times d_{13} &= 1 & \cdots (4)
 \end{aligned}$$

ここで、図2の乗算結果から、  
 $7 \times 6 = 1$

となることが分かるので、6×(4)式を計算し、未知

$$\begin{aligned}
 6 \times 7 \times d_{12} &= 6 \times 1 \\
 d_{12} &= 6 & \cdots (5)
 \end{aligned}$$

さらに、前記(5)式を(1)式或いは(2)式に設定して同様に未知データd13を求める。ここでの説明は、(2)式に(5)式を代入する方法により下記する

$$\begin{aligned}
 3 \times d_{12} \perp 4 \times d_{13} &= 5 & \cdots (2) \\
 3 \times 6 \perp 4 \times d_{13} &= 5 \\
 10 \perp 4 \times d_{13} &= 5 \\
 4 \times d_{13} &= 5 \perp 10 \\
 4 \times d_{13} &= 15 & \cdots (6)
 \end{aligned}$$

ここで、図2に示す乗算結果から、 $4 \times 13 = 1$ が分かるので、13×(6)式を計算し、未知データd13の係数を1とすることにより、d13の値を求めることができる。

$$\begin{aligned}
 13 \times 4 \times d_{13} &= 13 \times 15 \\
 d_{13} &= 7 & \cdots (7)
 \end{aligned}$$

データd12の係数を1とすることにより、d12の値を求める。

【0067】

(6)式を作成し、未知データd13を求める。

【0068】

できる。

【0069】

以上のような解法手順を踏むことにより、(5)式、(7)式によりそれぞれ未知データ $d_{12}=6$ 、 $d_{13}=7$ が求まり、始めに設定した仮定値と一致するので、未知データが正しく求められていることが分かり、障害ディスクのデータを復元されたことが分かる。

【0070】そこで、以下、前述と同様な手順に従って、2番目以降の全てのグループにおける障害ディスクのデータを復元化する。

【0071】なお、以上述べた具体例は、冗長するパリティの個数が2個の場合を例に上げたが、例えば冗長するパリティの個数が $k$ 個の場合、 $k$ 個の合同式を組み立てることができるので、 $k$ 個まで未知データを求めることができるので、 $k$ 個のデータが消失しても復元することができる。

【0072】さらに、上記実施の形態では、拡大ガロア体 $GF(2n)$ の $n$ を4として取り扱ったために、合同式のデータに付加する係数が1～15までの15個までしか扱えなかったのが、最大1グループ内のデータ数は、15以内であった。しかし、 $n$ をより大きな値とすることにより、より大きな値の係数まで扱うことができ、非常に大きなデータ数を扱うことができる。

【0073】さらに、RAID制御装置4としては、ハードウェアにより構成することもできる。すなわち、拡大ガロア体 $GF(2n)$ を用いて前述する要領によりデータ用ディスクの各グループごとに付加するパリティデータを算出する論理回路要素(ハードウェア)からなるパリティ算出手段と、任意のディスク障害時、その障害発生ディスクのデータを未知数として拡大ガロア体 $GF(2n)$ の規則に従って連立合同式を組み立て、この組み立てられた連立合同式を解くことにより、障害発生ディスクのデータを復元する論理回路要素(ハードウェア)からなるデータ復元手段とにより構成すれば、パリティ算出およびデータ復元の計算を高速度で行うことができる。

【0074】(その他の実施の形態)

(1) 図1に示す実施の形態では、データを記録するデータ用ディスクとは別に特定のパリティ用ディスクを設けた構成としたが、図3に示すように例えば5台のディスク7a～7eをそれぞれ6ブロックに分割し、各ディスク7a～7eの上位ブロックを1番目グループ、次のブロックを2番目グループ、以下、順次下位のブロックに対してグループ名を付けていく。その他の構成は図1と同様であるので省略する。

この実施の形態において特に異なるところは、5台のディスク7a～7eの各グループごとにそれぞれ任意の2つのブロックをパリティブロックとして設定することにある。

【0075】因みに、図3に示すディスク7a～7eの1番目グループではディスク7d、7eのブロックにパリティ $p_{14}$ 、 $p_{15}$ を設定し、2番目グループではデ

ィスク7c、7dのブロックにパリティ $p_{24}$ 、 $p_{25}$ を設定する。その他のグループでも同様にパリティを設定する。

【0076】その結果、1番目グループのデータ $d_{11}$ の一部書き換えたとき、それに伴ってパリティ $p_{14}$ 、 $p_{15}$ を変更する必要がある。以下、2番目以降グループのデータの一部を変更した場合でも、同様にパリティを変更することになる。

【0077】しかし、各グループごとに個別に異なるディスクにパリティを設定する構成とすれば、次のようなメリットがある。一般に、データの一部を変更した場合、幾つかのグループに対してまとめてパリティを変更しなければならない場合が多いが、各グループのパリティの格納位置が図3に示すごとく異なるディスクに格納する構成となっていれば、通常、パリティ書き換えのためにディスク要求が集中してディスク要求待ちとなるが、本実施の形態ではディスク要求が集中することがなく、ディスク要求待ちを解消できるメリットがある。

【0078】(2) 次に、本発明に係わる記録媒体の発明について説明する。

【0079】図4は記録媒体をもったRAID装置の全体構成を示す図である。

【0080】このRAID装置は、キーボード、マウスなどの入力装置11と、表示装置12と、後記するパリティ算出用およびデータ復元用プログラムを記録する記録媒体13と、この記録媒体13に記録されるパリティ算出用プログラムおよびデータ復元用プログラムの何れかを読み出して所定の機能を実現するCPUで構成されたデータ処理部14と、プログラムデータ、処理途中データ、処理結果のデータその他プログラム処理上必要なデータを一時記憶するデータバッファ15と、データ処理部14から導出されるバスラインに接続される複数のディスク161～16nによって構成されている。

【0081】なお、記録媒体としては、一般的には磁気ディスクが用いられるが、それ以外にも例えば磁気テープ、CD-ROM、DVD-ROM、フロッピーディスク、MO、CD-R、メモリカードなどを用いてもよい。

【0082】前記データ処理部14は、パリティ算出およびデータ復元に際し、次のような機能を実現する。

【0083】まず、パリティ算出処理にあっては、各データ用ディスクのグループ内データを順次読み出すデータ読み出し機能と、当該グループ内の各ブロックのデータに対し、拡大ガロア体 $GF(2n)$ に従って所定の係数を乗算するとともに、これら乗算値を加算するパリティ取得機能と、このパリティ取得機能によって取得したパリティを該当グループ内の該当ブロックに設定するパリティ設定機能と、上記一連の機能を、パリティ格納用ブロックおよび全グループについて繰り返し実行する機

能とを実現するものである。

【0084】一方、データ復元処理にあっては、ディスク故障時に正常なデータ用およびパリティ用ディスクのデータおよびパリティを読み出すデータ読み出し機能と、これら読み出したデータおよびパリティを用いてディスク故障台数に応じた連立合同式を作成し未知データを求めるデータ復元機能と、これら一連の機能についてデータ用ディスクのブロックおよび全グループについて繰り返してデータを復元する機能とを実現するものである。

【0085】次に、記録媒体13に記録されたプログラムの処理例について、パリティ算出処理およびデータ復元処理を参照して説明する。

【0086】(a) パリティ算出処理例について(図5参照)。

【0087】データ処理部14は、データディスクの各グループまたは全グループの各ブロックにデータを格納した後、例えば入力装置11からパリティ算出の指示を入力すると、記録媒体13からパリティ算出プログラムを読み出し、以下のような処理を実行する。

【0088】すなわち、データ処理部14は、データバッファ15などの不要データを消去する初期化処理を行った後、データバッファ15内カウンタメモリにディスクの1番目グループに相当するデータ「 $i=1$ 」をセットする。しかる後、各データ用ディスクの1番目グループに属する各ブロックのデータ読み出してデータバッファ15に記憶する(S1~S3、データ読み出し機能)。

【0089】さらに、データ処理部14は、拡大ガロア体GF(2n)に従い、読み出した1番目グループ内の各データに予め定める係数を乗算し、各データごとの乗算値データを求める。そして、これら各乗算値を加算することによりパリティを求める(S4、S5、パリティ取得機能)。

【0090】以上のようにしてパリティを求めたならば、パリティディスクの予め定める1つのブロックに設定する(S6、パリティ設定機能)。

【0091】しかる後、1番目グループ内の全パリティブロックへのパリティ設定完了かを判断し、未完了の場合にはステップS3に戻って同様の処理を実行する。一方、1番目グループ内のパリティ設定完了の場合には、全グループへのパリティ設定完了かを判断し(S8)、未だ完了していない場合にはカウンタメモリに+1をインクリメントし(S9)、ステップS3に戻り、同様の処理を繰り返して実行する。

【0092】(b) データ復元処理例について(図6参照)。

【0093】データ処理部14は、入力装置11からのデータ復元指示またはディスク故障時に自動的に立ち上がって記録媒体13からデータ復元プログラムを読み

取り、次のような処理を実行する。

【0094】すなわち、データ処理部14は、データバッファ15などの不要データを消去する初期化処理を行った後、データディスク故障を確認した後、データバッファ15内カウンタメモリにディスクの1番目グループに相当するデータ「 $i=1$ 」をセットする。しかる後、各データ用ディスクの1番目グループに属する各ブロックのデータ読み出してデータバッファ15に記憶する(S11~S14、データ読み出し機能)。

【0095】しかる後、データ処理部14は、失われたデータを未知データと正常なデータとを用いて、拡大ガロア体GF(2n)の規則に従い、未知データを左辺に置き、正常なデータを右辺に置き、データディスク故障台数に応じた連立合同式を作成し、この連立合同式を解くことにより未知データを求め、データバッファ15に格納する(S15、データ復元機能)。

【0096】さらに、1番目グループ内の故障データディスクのデータ全部復元かを判断し(S16)、当該グループの故障データ全部復元でないとき、既に求めた未知データを連立合同式に代入し、他の未知データを求める(S15、S16)。

【0097】さらに、ステップS16において当該グループ内故障データを全部復元したと判断したとき、全グループデータ復元完了かを判断し(S17)、未グループがあれば、カウンタメモリに+1をインクリメントした後(S18)、ステップS12に戻り、同様の処理を繰り返して実行し、全部の消失データを復元する。

【0098】従って、以上のような実施の形態によれば、データ処理部14は、記録媒体13に記録される図5のパリティ算出プログラム、図6に示すデータ復元プログラムを読み取って実行すれば、各ディスクにデータを格納後に容易、かつ、自動的にパリティを算出して設定でき、また複数台のデータディスクの故障時に失われたデータを迅速に復元させることができる。

【0099】なお、図4に示す記録媒体13のプログラムはアプリケーションソフトとして考えているが、OSの一部として使用してもよい。この場合には、記録媒体13に記録されるプログラムはデータ処理部14または上位コンピュータ6により読み出し可能とするものである。

【0100】(3) さらに、上記実施の形態では、冗長するパリティ算出の個数は各グループごとに2つのブロックと固定されていたが、この冗長するパリティ算出の個数をソフトウェアまたはハードウェア的に決定してもよい。

【0101】図7はかかる実施の形態例を示す構成図である。

【0102】このRAIDシステムは、伝送ライン20上にコンピュータを含むコントローラ21およびRAID装置22が接続されている。



【0103】このコンピュータを含むコントローラ21は、データ処理上必要なプログラムをRAID装置22にダウンロードしたり、当該RAID装置22に必要な指示、命令を与える機能をもっている。

【0104】前記RAID装置22は、図1や図3と同様なデータ配列構成をもつパリティを格納する複数のディスク231～23nと、プログラム処理開始、プログラム処理上必要な指示、さらには設定情報を入力するキーボードその他一般的に使用されている入力機器を含むハードウェアスイッチ24と、拡大ガロア体GF

(2n)を用いて前述する要領によりデータ用ディスクの各グループごとに付加するパリティデータを算出するソフトウェア的またはハードウェア的に構成されるパリティ算出手段および任意のディスク障害時、その障害発生ディスクのデータを未知数として拡大ガロア体GF

(2n)の規則に従って連立合同式を組み立て、この組み立てられた連立合同式を解くことにより、障害発生ディスクのデータを復元するソフトウェア的またはハードウェア的に構成されるデータ復元手段をもつRAID制御装置25とが設けられている。

【0105】ところで、以上のようなシステムにおいて、オペレータがコンピュータを含むコントローラ21側から冗長するパリティ算出の個数を入力し、RAID装置22にダウンロードするか、或いはオペレータがRAID装置22側のハードウェアスイッチ24からスイッチ操作によって冗長するパリティ算出の個数を入力する。

【0106】RAID制御装置25は、これら各グループごと或いは全グループに対して入力されたパリティ算出個数に従ってパリティを算出する。

【0107】従って、以上のような実施の形態によれば、オペレータがコントローラ21またはハードウェアスイッチ24から冗長するパリティ算出の個数を入力するようにすれば、用途、目的に応じて自在に冗長するパリティの個数を容易に変更でき、信頼性の向上および柔軟性を確保できる。

【0108】(4) 前述する実施の形態は、1台のRAID制御装置6、25が複数のディスク(2a～2

c、3a、3b)、(7a～7e)または(231～23n)を管理し、パリティ算出および失われたデータの復元化を実施する構成としたが、例えば図8に示すように複数のRAID制御装置25a、25b、25cがそれぞれ個別にディスクを管理する場合、例えばRAID制御装置25aがディスク23a1、23a2、RAID制御装置25bがディスク23b1～23b3、RAID制御装置25cがディスク23c1～23c3を管理するような構成のものがある。

【0109】このようなRAID装置22においては、複数のRAID制御装置25a～25c相互にデータ授受可能に連携させ、これらRAID制御装置25a～25c管理下のディスク23a1、23a2、23b1～23b3、23c1～23c3全体について、上位ブロックから順次ブロック26化し、これらグループ単位でデータdの書込みおよび読み出しの他、冗長するパリティpの算出やディスク故障により失われたデータの復元化処理を行うものである。

【0110】なお、パリティの算出およびデータの復元化は前述した通りである。

【0111】このような実施の形態によれば、各RAID制御装置ごとに必ずしも冗長するパリティのために1台分以上のパリティディスクを設置しなければならない制限がなくなり、RAID装置全体として冗長するパリティを低減化できる。

【0112】(5) 図1に示すRAID装置においては、各ディスクのデータ例えばd11、d12、d13を用いて、それぞれパリティp14、p15を個別に算出したが、例えば以下のような手段によりパリティを算出することにより、障害ディスク位置を把握し、データを修復するようにしてもよい。

【0113】具体的には、パリティp14、p15の算出するに際し、パリティp14を算出した後、パリティp15の算出に当たっては既に算出済みのパリティp14もデータの一部とみなして各データに異なる係数をかけて以下のように算出する。

【0114】

$$\begin{array}{rcl} d11 \perp & d12 \perp & d13 \\ d11 \perp & d12 \perp & d13 \perp & p14 \equiv 0 \end{array} \quad \dots (8)$$

$$2 \times d11 \perp 3 \times d12 \perp 4 \times d13 \perp 5 \times p14 \equiv p15 \quad \dots (9)$$

ここで、d11、d12、d13、p14、p15の何れか1つの情報に誤りe1が発生したとすると、p15に誤りが発生した場合を除き、前記(8)式の右辺に誤りe1が検出される。

【0115】さらに、前記(9)式においてパリティを計算すると、p15を除いて各データの係数に応じた誤りが検出されるので、データの係数を計算することによりd11、d12、d13、p14の何れかのディスクに誤りが発生している可能性があることを特定できる。

【0116】今、例えば各データd11、d12、d13、p14が以下の値のとき、データd13に誤りe1が発生したと想定して計算してみる。

【0117】d11≡11

d12≡12

d13≡13

e1 ≡ 8

mは各データに付けた異なる係数値とする。

【0118】パリティp14、p15の算出は、

$$\begin{array}{ccccccc} 11 \perp & 12 \perp & 13 & & \equiv p14 \equiv 10 \\ 2 \times 11 \perp 3 \times 12 \perp 4 \times 13 \perp 5 \times 10 & & \equiv p15 \equiv 7 \end{array}$$

ここで、d13にe1≡8のエラーが発生し、d13の値が13から5に変化したとする。

【0119】

$$\begin{array}{ccccccc} 11 \perp & 12 \perp & 5 \perp & 10 \equiv 8 \equiv & e1 \cdots (10) \\ 2 \times 11 \perp 3 \times 12 \perp 4 \times 5 \perp 5 \times 10 \equiv 1 \equiv m \times e1 \perp p15 \\ 5 \perp & 7 \perp & 7 \perp & 4 \equiv 1 \equiv m \times e1 \perp & 7 \cdots (11) \end{array}$$

この(10)式の両辺は、データに誤りが発生しなければ、前記(8)式で設定したように0とならなければならないが、誤りが発生した場合には0とならない。前記(11)式はデータに誤りがなければ、前記(9)式で設定したパリティp15と一致しなければならないが、誤りが発生した場合には各データの係数mに応じたm×

e1の誤りが含まれた状態で出てくる。

【0120】前記(10)式および(11)式により、mを計算し特定することにより、どのデータに誤りがあつたかを把握でき、さらに障害を起こしたディスクを特定することができる。

【0121】

$$\begin{array}{ccccccc} 8 \equiv & e1 & & & \cdots (10') \\ 1 \equiv m \times e1 \perp & 7 & & & \cdots (11') \\ \text{この(10')} & \text{式を(11')} & \text{式に代入し整理すると、} & & \\ 8 \times m \equiv & 7 \perp & 1 \equiv & 6 & \cdots (12) \end{array}$$

となる。図2の乗算結果図から、

$$8 \times 15 \equiv 1$$

となるように、(12)式の両辺に係数15を乗算するとともに、このmの係数を1とし、mの値を求める。

$$\text{【0122】 } 15 \times 8 \times m \equiv 15 \times 6$$

$$m \equiv 4$$

となり、d13の係数であることが分かる。ここで、さらに検算のために、4×(8)式⊥(9)式を用いて、データd13の要因を排除したところの下記(13)式を作成し検証する。

【0123】

$$\begin{array}{ccccccc} 4 \times d11 \perp & 4 \times d12 \perp & 4 \times d13 \perp & 4 \times p14 \equiv & 0 \\ 2 \times d11 \perp & 3 \times d12 \perp & 4 \times d13 \perp & 5 \times p14 \equiv & p15 \\ 6 \times d11 \perp & 7 \times d12 & & 1 \times p14 \equiv & p15 \\ & & & \cdots (13) \\ 6 \times 11 \perp & 7 \times 12 & & 1 \times 10 \equiv & 7 \\ 15 \perp & 2 & & 10 \equiv & 7 \\ & & & 7 \equiv & 7 \end{array}$$

前記(13)式において両辺の値が一致しないとき、始めに想定された1個所の誤りが否定されたこととなり、複数箇所のデータに誤りがあることを判断できる。同様の理由により、d13以外のd11、d12、p14についても、誤り箇所を特定できることが分かる。

【0124】また、パリティp15に1個所誤りが発生した場合、前記(13)式で作成されたデータd13の要因を排除した方法で各データそれぞれの要因を排除し、誤りがどのデータの要因を排除しても誤りが一定となる場合には、パリティp15自身に誤りがあると判断でき、一方、異なる誤りが算出された場合には2個所以上に誤りがあると判断できる。

【0125】誤り箇所が1箇所に特定できた場合、図1で説明した手順により容易に誤り箇所のデータを復元できる。

【0126】以上のようにして、冗長するパリティの個数が2つの場合、1個所のデータの復元が可能となる。一般に、冗長するパリティの個数が1つ増える毎にデータに付加されたと考えられる誤り要因を1つずつ排除できるので、

冗長するパリティ個数-1

のデータ訂正が可能なソフトウェアのRAID装置を提供できる。

【0127】従って、以上のような実施の形態によれば、ディスクに誤り検出機能がない場合でも容易に利用できる他、かかる誤り検出機能をもつことにより、より信頼性の高い装置を実現できる。

【0128】その他、本発明は、その要旨を逸脱しない範囲で、種々変形して実施することが可能である。

【0129】

【発明の効果】以上説明したように本発明によれば、拡大ガロア体GF(2n)を用いて、複数のディスクにデータを格納した後に迅速にパリティを算出でき、また同時に複数台のディスクが故障した場合でも、そのディスク内容を容易に復元することができる。

【0130】また、用途、目的に応じて、冗長するパリティの個数を任意に変更可能であり、信頼性の向上および柔軟性に富んだ装置を実現できる。

【0131】さらに、複数のRAID制御装置管理下にある各ディスク全体をグループ化し、各グループごとに

冗長するパリティを設定するので、必ずしも冗長するパリティのために1台分以上のパリティディスクを設置しなければならない制限がなくなり、RAID装置全体として冗長するパリティを低減化できる。

【0132】さらに、パリティの算出から障害ディスクを容易に見つけ出し、データの修復を行うことができる。

【0133】さらに、本発明の記録媒体においては、パリティの算出および同時に複数台のディスク故障時にそのディスク内容を正確に復元可能なプログラムを記録した記録媒体を提供できる。

【図面の簡単な説明】

【図1】 本発明に係わるRAID装置の一実施の形態を示す構成図。

【図2】 拡大ガロア体GF(2<sup>n</sup>)による乗算結果を説明する図。

【図3】 本発明に係わるRAID装置の他の実施の形態を説明するものであって、グループ単位ごとに任意のブロックにパリティを設定する図。

【図4】 本発明に係わる記録媒体を説明するコンピュータシステムの構成図。

【図5】 図4に示す記録媒体に記録されるプログラムの処理例を説明するフローチャート。

【図6】 図4に示す記録媒体に記録されるプログラムの他の処理例を説明するフローチャート。

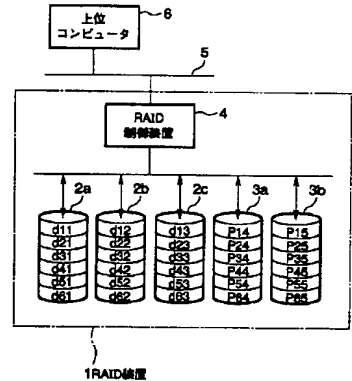
【図7】 本発明に係わるRAID装置の他の実施の形態を示す構成図。

【図8】 本発明に係わるRAID装置のさらに他の実施の形態を示す構成図。

【符号の説明】

- 1, 22…RAID装置
- 2a, 2b, 2c…データ用ディスク
- 3a, 3b…パリティ用ディスク
- 4, 25a, 25b, 25c…RAID制御装置
- 6…上位コンピュータ
- 7a~7e, 161~16n, 23a1, 23a2, 23b1~23b3, 23c1~23c3…ディスク
- 13…記録媒体
- 21…コンピュータを含むコントローラ
- 22…記録媒体
- 26…グループ

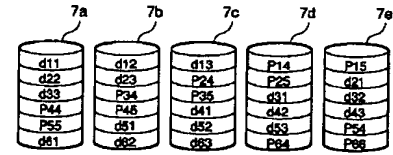
【図1】



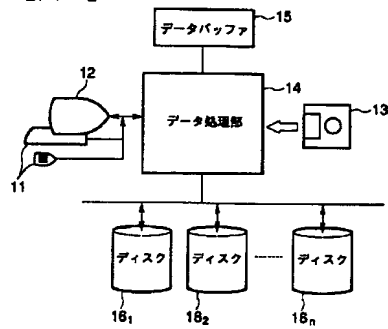
【図2】

拡大ガロア体GF(2 <sup>4</sup> )の乗算結果																
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	0	2	4	6	8	10	12	14	3	1	7	5	11	9	15	13
3	0	3	6	5	12	15	10	9	11	8	13	14	7	4	1	2
4	0	4	8	12	3	7	11	15	6	2	14	10	5	1	13	9
5	0	5	10	15	7	2	13	8	14	11	4	1	9	12	3	6
6	0	6	12	10	11	13	7	1	5	3	9	15	14	8	2	4
7	0	7	14	9	15	8	1	6	13	10	3	4	2	5	12	11
8	0	8	3	11	6	14	5	13	12	4	15	7	10	2	9	1
9	0	9	1	8	2	11	3	10	4	13	5	12	6	15	7	14
10	0	10	7	13	14	4	9	3	15	5	8	2	1	11	6	12
11	0	11	5	14	10	1	15	4	7	12	2	9	13	6	8	3
12	0	12	11	7	5	9	14	2	10	6	1	13	15	3	4	8
13	0	13	9	4	1	12	8	5	2	15	11	6	3	14	10	7
14	0	14	15	1	13	3	2	12	9	7	6	8	4	10	11	5
15	0	15	13	2	9	6	4	11	1	14	12	3	8	7	5	10

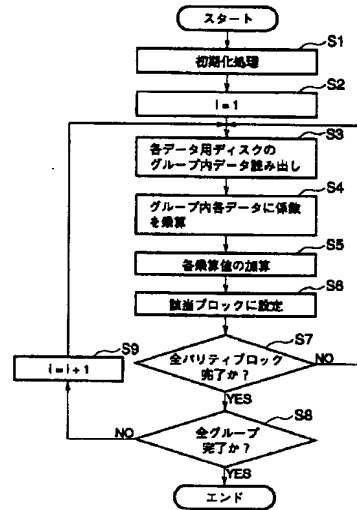
【図3】



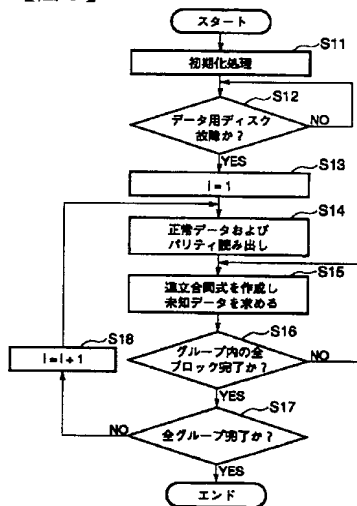
【図4】



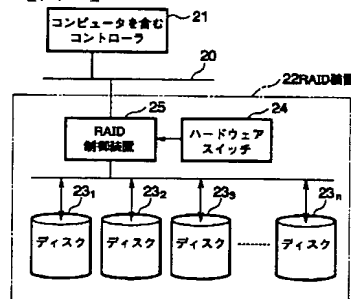
【図5】



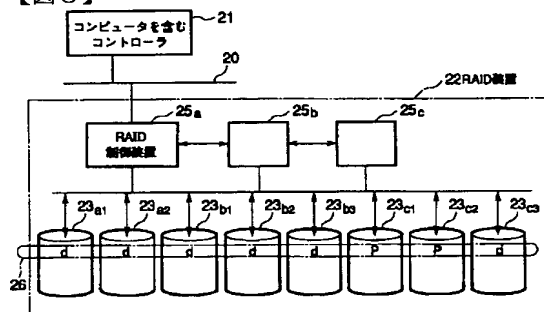
【図6】



【図7】



【図8】



フロントページの続き

(51) Int. Cl. 7	識別記号	F I	テーマコード* (参考)
G 1 1 B 19/04	5 0 1	G 1 1 B 19/04	5 0 1 G
			5 0 1 D
20/18	5 7 0	20/18	5 7 0 Z